Zhang B. and Zhu Z. (2017). "Vision-based detection of fall accident at flat level surfaces". In: *LC3 2017: Volume I – Proceedings of the Joint Conference on Computing in Construction (JC3)*, July 4-7, 2017, Heraklion, Greece, pp. 177-184. DOI: https://doi.org/10.24928/JC3-2017/0198.

# VISION-BASED DETECTION OF FALLS AT FLAT LEVEL SURFACES

**Bingfei Zhang[1] and Zhenhua Zhu[2]**

**Abstract:** Workers might experience fall accidents even when they are working at flat level surfaces. These accidents plus other types of fall accidents have been reported as one of the major causes for worker-related fatalities and injuries. Currently, it becomes common to set up video cameras to monitor working environments. The video cameras provide an alternative to detect fall accidents. The objective of this paper is to investigate the feasibility of detecting fall accidents of workers with video. The preliminary focus is put on the fall detection under one single monocular camera. A novel fall detection method is proposed. Under the method, workers in the videos captured by the video cameras are first detected and tracked. Their pose and shape related features are then extracted. Given a set of features, an artificial neural network (ANN) classifier is further trained to automatically determine whether a fall happens. The method has been tested and the detection precision and recall were used to evaluate the method. The test results with high detection precision and recall indicated the method effectiveness. Also, the lessons and findings from this research are expected to build a solid foundation to create a vision-based fall detection solution for safety engineers.

**Keywords:** Fall detection, video processing, computer vision, safety management.

## 1 INTRODUCTION

Overall, fall accidents are considered as common hazards and major causes for serious injuries. Every year, the U.S. Bureau of Labour Statistics reports hundreds of fatalities as well as tens of thousands of nonfatal injuries due to falls (U.S. BLS 2012). Also, 15% accidental deaths were caused by fall accidents, which were ranked as the second major cause (Kendzior 2010). In addition to deaths and injuries, the financial loss associated with the fall accidents is also tremendous. It was estimated that the total cost of unintentional fall accidents in the public and private sectors could be $80 million each year in U.S. (Kendzior 2010).

In order to reduce the loss incurred due to the fall accidents, it is important to detect them promptly once the accidents happen. This way, the rescue actions for the victims in the accidents could be initiated immediately. Also, the corresponding emergency response plan could be arranged in a timely manner. So far, one common category of the methods for the fall detection mainly relies on the use of wearable sensor, such as accelerometers, barometric pressure sensors, gyroscopic sensors or combination of them (Makantasis et al. 2015). However, it is always required to attach these wearable sensors on the human body, in order to obtain the useful information for the fall detection. As a result, the popularity of the use of these methods in practice is limited.

[1] Department of Building, Civil, and Environment Engineering, Concordia University, Montreal, Canada, H3G 1M8; email: bingfeizh@gmail.com

[2] Department of Building, Civil, and Environment Engineering, Concordia University, Montreal, Canada, H3G 1M8; email: zhenhua.zhu@concordia.ca

Another wide category of research studies focuses on the fall detection with computer vision techniques. In those research studies, video cameras are adopted to capture fall accidents when they happen. It becomes common to set up video cameras to monitor working environments, due to the recent fast development of digital camera technology. Thus detecting fall with video cameras become an alternative solution. However, there are still challenges on the video-based fall detection. These challenges include but are not limited to the occlusions, illumination variations, and clutters, all of which have significant impacts on the fall detection precision and recall.

This paper proposed a video-based fall detection method with one single camera. It is the first step to detect fall accidents with digital camera technology. Under the method, workers are first detected and tracked in the videos using existing detection and tracking algorithms. Based on the detection and tracking results, the workers' pose and shape related features are extracted. These features are input into an artificial neural network (ANN) classifier to automatically determine whether a fall accident happens. The classifier was pre-created through the supervised training with a set of training samples. So far, the method has been tested in the laboratory environment. The detection precision and recall were used to evaluate the method effectiveness. The test results showed the method could achieve both high precision and recall. The lessons and findings learned are expected to build a solid foundation to help safety engineers timely rescue the victims in the fall accidents.

## 2 Background

Many vision-based fall detection methods have been proposed in recent years. These methods adopt different camera systems, including multi-camera systems, monocular camera systems, and depth camera systems (Sathyanarayana et al. 2015). Multi-camera systems are mainly used to acquire the 3D features of fall postures for the purpose of the detection. A 3D shape of a person could be generated by the multi-camera systems, and the distribution of the 3D shape is the used to decide whether a fall accident happens or not (Auvient et al. 2011). Apart from the 3D shape, the principal component and variance ratio of the 3D human silhouette are also calculated from multi-view images and used for detect fall accidents (Hazelhoff et al. 2008). In most multi-camera systems, the features extracted or processed from each single camera are combined with a fusion unit, so that these features could be complementary with each other to conduct the fall detection (Sathyanarayana et al. 2015). The advantage is that the detailed 3D information for the fall detection could be acquired from the multiple camera views. However, the accurate calibration and synchronized video sequences are needed in order to get the reliable data. Also, it might be difficult to guarantee the real-time requirement with the affordable camera hardware configurations.

Monocular camera systems are also used for fall detection. The systems, unlike the multi-camera systems, focus on the 2D features for the fall detection. These features, for example, include but are not limited to the height-width ratio of bounding box, the velocity of the centre of bounding box and the angle of bounding ellipse (Foroughi et al. 2008). The problem of the 2D features is that the distance between the camera and the person would influence the reliability of the extracted features. In order to address such a problem, several methods for generating the 3D feature with a monocular camera system are proposed. In doing so, the camera calibration and inverse perspective mapping are used (Makantasis et al. 2012).

In order to get the 3D features for the fall detection with one single camera, the idea of using the depth camera is also proposed. The depth camera utilized the time-of-flight principle. This way, the actual vertical velocity (Mastorakis and Makris 2012) and 3D motion history (Dubey et al. 2012) obtained by the depth camera could be used for the fall detection. Although the depth camera is able to get the 3D information easily and fast, the camera is usually equipped with short-range sensors. Therefore, it is not capable of providing a wide field of view and monitoring a large area.

## 3 Objective

This paper proposed a method to detect fall accident on flat level. The proposed method utilized both 2D and 3D information with one single camera. A neural network is used to make the detection more accurate. Compared with other monocular camera fall detection method, the proposed method generated a more robust result with 3D information. Moreover, compared with the multi-camera systems, the proposed method is less constrained to the infrastructure. The proposed method was tested in the laboratory environment, and the experiment results proved that the proposed method is efficient.

## 4 Proposed methodology

### 4.1 Feature extraction

Fall incidents could be described by motion features, including the height-width ratio, the angle of bounding ellipse and the vertical velocity. When a person is falling, the height-width ratio of his or her bounding box is always smaller than the one when he or she is standing. In addition, the angle of the bounding ellipse is close to 90° when the person is standing and close to 0° after the fall. The vertical motion velocity could also reflect the motion of the person. The value of the velocity would be larger when a fall accident occurs. Thus all these three features are used for detecting a fall.

In order to extract all these features, first of all, the persons should be extracted from the images. A foreground extraction algorithm (Duolamis et al. 2010) was used. The algorithm uses the "pyramidal" Lucas-Kanade algorithm to estimate the intensities and the directions of the motion vectors in the scene. With the foreground extraction algorithm, the persons are extracted and the images are transformed into black-white format, as shown in Figure 1. At the same time, the minimum bounding boxes are calculated. The four corners (top-right, top-left, bottom-right and bottom-left) of a bounding box are recorded as ptr (pt, pr), ptl (pt, pl), pbr (pb, pr), pbl (pb, pl).

Figure 1: Extracted foreground

The first feature to be calculated is the height-width ratio. The height-width ratio is defined as the ration between the height and the width of the bounding box. As the

coordinates of the four corners of the bounding box are known, the ratio could be expressed by the following equation:

$$R = \frac{h}{w} = \frac{p_t - p_b}{p_r - p_i} \quad (1)$$

where h is the height of the bounding box and w is the width of the bounding box. For a standing person, the ratio is higher than a fell down person.

In order to obtain the orientation of bounding ellipses of the persons, the image moments which could describe the extracted foreground were used. The image moment Mij of a scalar image is defined as:

$$M_{ij} = \sum_x \sum_y x^i y^j I(x,y) \quad \text{for i, j} = 0, 1, 2 \ldots \quad (2)$$

where I(x, y) is the pixel intensity at the point (x, y).

The orientation of the ellipse could be obtained by the central moments of second order. The central moment μij is given by

$$\mu_{ij} = \sum_x \sum_y (x - x_c)^i (y - y_c)^j I(x,y) \quad \text{for i, j} = 0, 1, 2 \ldots \quad (3)$$

and the orientation θ is obtained by

$$\theta = \frac{1}{2} \arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right) \quad (4)$$

Also, the center and the semi-axis of the ellipse could be calculated with image moments. The bounding ellipses are drawn with the three features calculated above and shown in Figure 2.

Figure 2: Bounding box and bounding ellipse

Instead of using person's projection height, the person's actual height is used in this method. This is because the actual height won't be influenced by the distance between the person and the camera. In addition, the actual height could give an information about the type of the moving object extracted from the back ground, then moving machines or pets won't be regarded as a person. What's more, with the actual height, the position and view of the camera would not be restricted. In order to obtain the actual height of the person, camera calibration is needed to be done at first.

Tsai's calibration algorithm (Tsai 1987) is used in this method. The world coordinate of a point and the image coordinate of its projection point on the image are corresponded. The internal and external parameters are calculated by these corresponding points. Traditional methods use chessboards to do the calibration. However, due to the large scale of the sites, the chessboard should be extremely large to be seen clearly on the images. What's more, to make the calibration accurate, the corresponding points should be distributed evenly on the whole image. It's obviously that the chessboard is inappropriate. In this method, a laser scanner was used to obtain the world coordinates of the points. The accuracy of the laser scanner is ±1mm which is accurate enough for the calibration. After calibration, the internal parameters including the focal length f, the skew factor sx and the

distortion k, as well as the external parameters including the rotation matrix R and the translation matrix T could be obtained.

In this method, the site is assumed as a plane, thus the persons' feet are on the same plane. The left-bottom point of the bounding box is considered as the foot of a person thus the person's position on the plane could be determined. Through the image coordinate of the left-bottom point, the projection line could be determined with the calibration result, and the intersection of the projection line and the working plane is deemed to be the position of the person on the plane. By this way the world coordinate of the person's foot (xww, yww) is determined. As the horizontal coordinates of the camera (xwc, ywc) could be generated from the laser scanning result, the distance Z between the person and the camera could be obtained by

$$Z = \sqrt{(x_{ww} - x_{wc})^2 + (y_{ww} - y_{wc})^2} \tag{5}$$

As we are using a pinhole camera model, the actual height H is proportional to the image height h, and the relationship could be described as

$$H = \mathrm{Z}\frac{h}{f} \tag{6}$$

where f is the focal length of the camera. The relationship is shown in the Figure 3. As h is calculated above and f could be obtained from the calibration, the actual height of the person could be obtained.

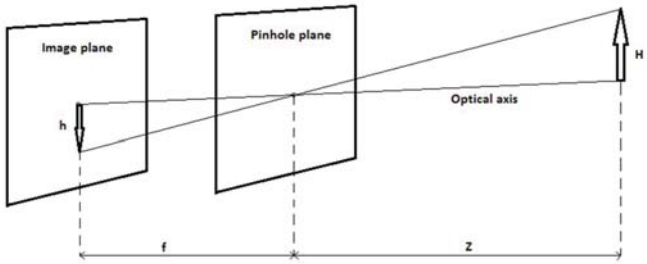


Figure 3: Pinhole model

## 4.2 Artificial neural network for fall classification

After obtained the three features of persons' motion, a Back Propagation Neural Network is trained to detect the fall. In order to avoid falling to a local minimum, a Genetic Algorithm is used to optimize the initial weight and threshold values. Once the training of the neural network is done, the algorithm could detect fall from video automatically.

The training process is finished in the lab. The inputs are a 4x1 vectors. The former three elements are based on the three features extracted above and the fourth element is a label for distinguishing fall. The video is extracted into frames of k fps, and then the difference between the three features extracted from two successive frames are regarded as the former three elements. In this case, k is set to be 5, which means the time between two images is 200ms. This time is sufficient for detecting a fall and discriminating it than other activities. For the last element, the value will be set as 1 if a fall appears and the value will be 0 oppositely.

The neural network would work after the training. The input of the network is the difference of the three features between two successive frames and the output is the label 0 or 1, 1 for fall and 0 for not fall. Considering the precision, only if 5 in 10 frames are labelled with 1, a fall will be alerted by the algorithm. The time window of 5 frames is about 1s which is the average duration of a fall incident.

## 5 EXPERIMENT RESULTS AND DISCUSSION

The proposed method was tested in the lab. A GoPro Hero 4 was fixed on the top of wall to simulate a surveillance camera. (See Figure 4.) A video was taken, in which one person did several activities including fall, sit on floor, and lie down on floor. (See Figure 5.) The accuracy of the proposed method was evaluated.

Figure 4: Set of camera

Figure 5: (From left to right) lie down, stand, fall and sit

Figure 6 illustrates some example results. The green bounding box means that no fall was detected and the red bounding box means that a fall was detected.

Figure 6: Fall detection results

In order to evaluate the accuracy of the proposed method, the precision and the recall are defined as follows:

$$Precision = TP/(TP + FP) \quad (7)$$

$$Recall = TP/(TP + TN) \quad (8)$$

where TP means true positive, FP means false positive and TN means true negative. The result of the proposed method is shown in the Table.1.

Table 1: Result of fall detection

| Actions | Number | Precision | Recall |
|---|---|---|---|
| Fall | 10 | 100% | 62.5% |

This method had a high precision rate and a low recall rate. Considering the hazard results of the fall incidents, precision was more important than recall. Thus, the proposed method might be suitable for fall detection. However, the low recall value was also a problem. That problem may be caused by the similarity of the posture of fall and other activities. In order to increase the recall rate, the threshold of fall could be changed. If a fall accident was defined when 6 in 10 frames were labelled as a fall, the precision would be reduced slightly but the recall would increase significantly. The result is shown in Table 2.

Table 2: Precision and recall rate according to fall frames number threshold

| Thresholds | Precision | Recall |
|---|---|---|
| 5 of 10 frames | 100% | 62.5% |
| 6 of 10 frames | 90% | 83% |

The selection of the threshold should be decided by the practical circumstance. If the focus is on procession rate, the thresholds should be set higher, thus although some non-fall will be detected as a fall, no fall accident will be omitted. Oppositely, if the recall rate is more important, the thresholds should be set lower, thus no non-fall will be detected as fall, then the needless reactions are avoided.

Additionally, the training procedure could significantly affect the detection results. The features extracted from a fast sitting/lying down person is similar to a falling person, which will mislead the neural network. To avoid such situation, proper numbers of training frames are recommended. Less frames will lead to under fitting and more frames will lead to over fitting. In this experiment, 300 frames were used for training (100 positives and 200 negatives) to reach the highest precision and recall rate.

## 6 Conclusion

Fall accidents are one of the leading causes for serious work related fatalities. In order to alleviate the loss brought by the fall accidents, it is necessary to detect them in a timely manner, once the accidents happen. This paper proposed a video-based method for the automatic detection of fall accidents. The method utilized the 3D features estimated from a monocular camera system and created an ANN classifier. The proposed method has been tested and it was found that the detection precision and recall could reach up to 90% and 83%. Compared with the method proposed by Liu et al., which average accuracy is 90.68%, the proposed method is acceptable. Future work will be focused on investigating the feasibility of using the deep learning techniques to increase the fall detection precision and recall ratios.

## 7 REFERENCES

Auvinet, E., Multon, F., Saint-Arnaud, A., Rousseau, J. and Meunier, J. (2011). Fall Detection with Multiple Cameras: An Occlusion-Resistant Method Based on 3-D Silhouette Vertical Distribution. *IEEE Transactions on Information Technology in Biomedicine*, 15(2), pp. 290-300.

Doulamis, N. (2010). Iterative motion estimation constrained by time and shape for detecting persons' falls. In: Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments, ACM, New York, pp. 62:1-62:8.

Dubey, R., Ni, B., Moulin, P. (2012). A depth camera based fall recognition system for the elderly. In: International Conference Image Analysis and Recognition, Springer, Berlin Heidelberg, pp. 106–113.

Foroughi, H., Aski, B.S., Pourreza, H. (2008). Intelligent video surveillance for monitoring fall detection of elderly in home environments. In: 11th international conference on computer and information technology, IEEE, pp. 219–224.

Hazelhoff, L., Han, J., With PH (2008). Video-Based Fall Detection in the Home Using Principal Component Analysis. In: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer, Berlin Heidelberg, pp. 298–309.

Liu, M., Hong, D., Han, S., Lee, S. (2016). Silhouette-Based On-Site Human Action Recognition in Single-View Video. In: Construction Research Congress 2016, American Society of Civil Engineers (ASCE), pp. 951-959.

Kendzior, R. (2010). *Falls aren't funny*. Lanham, Md.: Government Institutes.

Makantasis, K., Protopapadakis, E., Doulamis, A., Grammatikopoulos, L., Stentoumis, C. (2012). Monocular Camera Fall Detection System Exploiting 3d Measures: A Semi-Supervised Learning Approach. In: European Conference on Computer Vision Springer, Berlin Heidelberg, pp. 81–90.

Makantasis, K., Protopapadakis, E., Doulamis, A., Doulamis, N., & Matsatsinis, N. (2015). 3D measures exploitation for a monocular semi-supervised fall detection system, *Multimedia Tools and Applications*, 75(22), pp. 1-33.

Mastorakis, G. and Makris, D. (2012). Fall detection system using Kinect's infrared sensor. *Journal of Real-Time Image Processing*, 9(4), pp.635-646.

Sathyanarayana, S., Satzoda, R., Sathyanarayana, S. and Thambipillai, S. (2015). Vision-based patient monitoring: a comprehensive review of algorithms and technologies. *Journal of Ambient Intelligence and Humanized Computing*.

Tsai, R.Y. (1987). A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses. IEEE Journal of Robotics and Automation, RA-3(4), pp. 323-344.

U.S. Bureau of Labour Statistics (BLS), *Occupational injuries/illnesses and fatal injuries profiles*, http://data.bls.gov/gqt.